UNIVERSIDADE FEDERAL DE UBERLÂNDIA
Faculdade de Computação
Av. João Naves de Ávila, nº 2121, Bloco 1A - Bairro Santa Mônica, Uberlândia-MG, CEP 38400-902
Telefone: (34) 3239-4144 - http://www.portal.facom.ufu.br/ facom@ufu.br

PLANO DE ENSINO

1. IDENTIFICAÇÃO

Componente Curricular:	Organização e Recuperação da Informação							
Unidade Ofertante:	Faculdade de Computação							
Código:	GSI024		Período/Série:		5	Turma:	S	
Carga Horária:						Natureza:		
Teórica:	30	Prática:	30	Total:	60	Obrigatória:	(X)	Optativa: ( )
Professor(A):	Rodrigo Sanches Miani					Ano/Semestre:	2022/1	
Observações:								

2. EMENTA

Conceitos de documento, palavra e termo. Indexação de documentos: extração de termos, *stopwords*, *stemming*, criação de índices. *Thesauri*. Modelos de processamento de consultas. Avaliação de Sistemas de Recuperação de Informação (RI). RI em Documentos semi-estruturados, multimídia e documentos na Web. Extração da informação. Classificação de documentos. Redução de dimensionalidade.

3. JUSTIFICATIVA

A Organização e Recuperação de Informação (ORI) é uma disciplina abrangente da Ciência da Computação que se concentra principalmente em prover aos usuários o acesso fácil às informações de seu interesse. Em particular, essa disciplina trata da representação, armazenamento, organização, e acesso a itens de informação, como documentos, páginas da Internet, catálogos online, registros estruturados e semiestruturados, objetos multimídia e etc. Devido ao volume gigantesco de informação gerado pelos sistemas de informação, as máquinas de busca de informação se tornaram ferramentas fundamentais para localizar e recuperar informação. Portanto, os conhecimentos ligados ao funcionamento e construção de máquinas de busca são indispensáveis para permitir uma formação atualizada e multidisciplinar do bacharel em Sistemas de Informação.

4. OBJETIVO

Objetivo Geral:

Propor soluções para o problema de recuperar informações nos documentos de uma determinada coleção (estruturada ou semi-estruturada), a partir de uma consulta formulada pelo próprio usuário.

Objetivos Específicos:

Dominar os conceitos dos três modelos clássicos de RI (booleano, vetorial e probabilístico), assim como as técnicas frequentemente utilizadas para a construção de uma máquina de busca como indexação de documentos usando índices invertidos e o pré-processamento de documentos (análise léxica, eliminação de *stopwords*, stemming e seleção de palavras-chave).

5. PROGRAMA DA DISCIPLINA

1. Introdução à Recuperação da Informação e modelo booleano
2. Dicionário e lista de postings: conceitos de documento, palavra e termo.
3. Indexação de documentos: termos, *stopwords*, *stemming*, Thesauri
4. Compressão de índices
5. Peso de termos
6. Modelo Vetorial
7. Avaliação de sistemas de recuperação de informação
8. Realimentação de relevantes e expansão de consultas
9. Recuperação em documentos semi estruturados (XML)
10. Modelo probabilístico
11. Classificação de documentos
12. Agrupamento de documentos
13. Redução de dimensionalidade
14. Web: busca, *crawling*, indexação, análise de *links*
15. Extração da informação
16. Introdução à recuperação de imagens baseada em conteúdo

6. METODOLOGIA

- Aulas expositivas (quadro e *datashow*).

- Aulas práticas em laboratório, com atividades individuais.

- Atividades assíncronas complementares.

a) Atividades presenciais teóricas: 30 horas/aula

Horários das atividades presenciais teóricas: Terças, 19h00-20h40

b) Atividades presenciais práticas: 30 horas/aula

Horários das atividades presenciais práticas: Quintas, 20h50-22h30

c) Atividades assíncronas (Art 1º da Resolução CONSUN nº 30/2022): 12 horas/aula

Descrição da realização: desenvolvimento de estudos dirigidos propostos pelo docente ao longo da disciplina.

Plataforma de T.I. /softwares que serão utilizados: Para o desenvolvimento das atividades assícronas, serão utilizados softwares de edição de texto (*MS Office*, *Open Office* ou LaTeX-overleaf.com) a critério do discente e a plataforma virtual *MS Teams* para a entrega das atividades ao docente de acordo com o descrito no CRONOGRAMA DAS ATIVIDADES AVALIATIVAS apresentado em 8. AVALIAÇÃO.

Endereço web de localização dos arquivos: Arquivos a serem disponibilizados no *MS Teams*.

d) Demais atividades letivas: 0 horas;

Semana	Data	Conteúdo	QP	Carga horária presencial	Carga horária assínc
1	27/set	Não haverá aula - afastamento para visita técnica - Projeto CAPES/STIC/AMSUD			
1	29/set	Não haverá aula - afastamento para visita técnica - Projeto CAPES/STIC/AMSUD			
2	04/out	Não haverá aula - afastamento para visita técnica - Projeto CAPES/STIC/AMSUD			
2	06/out	Não haverá aula - afastamento para visita técnica - Projeto CAPES/STIC/AMSUD			
3	11/out	Apresentação do curso		2	
3	13/out	Introdução aos sistemas de recuperação de informação - Parte 1		2	
4	18/out	Introdução aos sistemas de recuperação de informação - Parte 2		2	
4	20/out	Elementos básicos dos sistemas de recuperação de informação - Parte 1		2	
5	25/out	Elementos básicos dos sistemas de recuperação de informação - Parte 2		2	
5	27/out	Elementos básicos dos sistemas de recuperação de informação - Laboratório		2	2
6	01/nov	Modelos de recuperação da informação - Modelo booleano		2	
6	03/nov	Modelos de recuperação da informação - Modelo vetorial	QP1	2	
7	08/nov	Modelos de recuperação da informação - Modelo vetorial		2	
7	10/nov	Modelos de recuperação da informação - Modelo probabilístico		2	
8	15/nov	Não haverá aula - Feriado			
8	17/nov	Modelos de recuperação da informação - Modelo probabilístico		2	2
9	22/nov	Modelos de recuperação da informação - Laboratório		2	
9	24/nov	Avaliação da recuperação da informação - Parte 1		2	
10	29/nov	Avaliação da recuperação da informação - Parte 2		2	
10	01/dez	Avaliação da recuperação da informação - Laboratório	QP2	2	2
11	06/dez	Realimentação de relevância - Parte 1		2	
11	08/dez	Realimentação de relevância - Parte 2	QP3	2	
12	13/dez	Realimentação de relevância - Parte 3		2	
12	15/dez	Realimentação de relevância - Laboratório		2	2
13	20/dez	Análise de Links - Parte 1		2	
13	22/dez	Análise de Links - Parte 2	QP4	2	
14	03/jan	Não haverá aula - Recesso			
14	05/jan	Análise de Links - Laboratório		2	2
15	10/jan	Outras aplicações - Classificação/Agrupamento de documentos		2	
15	12/jan	Outras aplicações - Classificação/Agrupamento de documentos	QP5	2	
16	17/jan	Outras aplicações - Classificação/Agrupamento de documentos		2	
16	19/jan	Outras aplicações - Classificação/Agrupamento de documentos - Laboratório		2	2
17	24/jan	Outras aplicações - Análise de sentimentos		2	
17	26/jan	Outras aplicações - Análise de sentimentos - Laboratório		2	
18	31/jan	Encerramento do curso - discussão sobre a disciplina		2	
18	02/fev	Recuperação		2	
		Carga horária síncrona total (hora-aula) = 60			
		Carga horária assíncrona total (hora-aula) = 12			
		Carga horária total (síncrona + assíncrona) = 72			

7. ATENDIMENTO E COMUNICAÇÃO COM OS DISCENTES

O atendimento aos alunos ocorrerá toda quinta-feira das 18:00 até 19:00 na sala 1B148.

A comunicação assíncrona com a turma será por meio de mensagens no *Microsoft Teams*.

8. AVALIAÇÃO

Os alunos serão avaliados de duas formas: provas (quiz presenciais - QPs) e trabalhos práticos (TPs). Teremos cinco (5) QPs e sete (7) TPs.

QP - são testes para avaliar se os alunos estão absorvendo o conteúdo. São testes curtos, objetivos (múltipla escolha) e associados a determinados assuntos vistos no curso. QP1 diz respeito a avaliação dos tópicos 1 e 2. QP2 do tópico 3, QP3 do tópico 4, QP4 do tópico 5 e QP5 do tópico 6.

TPs - são trabalhos que começarão a ser desenvolvidos em sala de aula e devem ser finalizados de forma assíncrona. Em geral, envolverão questionários e resolução de problemas práticos.

NF = 45*QP + 55*TP

Importante: entregas com atraso serão penalizadas. 1 dia de atraso = dedução de 10%, 2 dias de atraso = dedução de 15% porcento, 3 dias de atraso = dedução de 20% porcento, 4 dias de atraso ou mais = dedução de 40% porcento.

O discente terá direito a **Atividade de recuperação de aprendizagem**, se e somente se, **não obtiver o rendimento mínimo para aprovação e com frequência mínima de 75% (setenta e cinco por cento),** de acordo com o Art 141. das novas Normas Gerais de Graduação (Resolução CONGRAD Nº 46/2022). A frequência será aferida por chamada feita em sala de aula. Nesse caso, ele poderá fazer uma prova sobre o conteúdo de toda a disciplina;

O aluno será aprovado caso (NA+NR)/2 >= 60, onde NA = Nota atual do aluno e NR = Nota da recuperação do aluno;

Em caso de recuperação, a nota máxima do aluno será de 60 pontos.

9. **BIBLIOGRAFIA**

Básica

BAEZA-YATES, R.; RIBEIRO NETO, B. Recuperação de Informação: Conceitos e Tecnologia das Máquinas de Busca [S. l]: Bookman, 2013.

BAEZA-YATES, R.; RIBEIRO NETO, B. Modern information retrieval. 2. ed. São Paulo: Addison-Welsey, 2011.

MANNING, C.; RAGHAVAN, P.; SCHÜTZE, H. An introduction to information retrieval. Cambridge: Cambridge University Press, 2009. Disponível em https://nlp.stanford.edu/IR-book/information-retrieval-book.html

Complementar

CRESTANI, F.; PASI, G. Soft computing in information retrieval**:** techniques and applications. New York: Springer Verlag, 2010.

CROFT, B.; METZLER, D.; STROHMAN, T. Search engines**:** information retrieval in practice. São Paulo: Addison Wesley, 2009.

FRAKES, W. B.; BAEZA-YATES, R. Information retrieval & data structures. New Jersey: Prentice Hall, 1992.

MOENS, M. F. Information extraction: algorithms and prospects in a retrieval context. New York: Springer Verlag, 2006.

SUMMERFIELD, M. Programação em Python 3. Alta Books, 2013.

MELO, W. Introdução ao Universo da Programação com Python, 2021. Disponível em *https://wendelmelo.net/book*

10. **APROVAÇÃO**

Aprovado em reunião do Colegiado realizada em: ____/____/_____

Coordenação do Curso de Graduação: _____